Network Working Group
Internet-Draft
Intended status: Informational
Expires: September 14, 2017

L. Geng
China Mobile
S. Bryant
J. Dong
Huawei Technologies

March 13, 2017

Network Slicing Architecture
draft-geng-netslices-architecture-00

Abstract

This document defines the overall architecture of network slicing. Base on the general architecture, basic concepts of network slicing and examples of network slicing instances are introduced for clarification purposes. Some architectural considerations about the data plane, control plane, management and orchestration of network slicing are described to give a general view of network slicing implementation principles. This also helps to identify the gaps in existing IETF works relating to network slicing.

Status of This Memo

Copyright Notice

Table of Contents

1. Introduction

The Internet has always been designed to support a variety of services. The emerging 5G market is expected to bring this diversity of services to a new level. Typical examples of new bandwidth-hungry services enabled by 5G include high definition (HD) video, virtual reality (VR) and augmented reality (AR). The high bandwidth requirement of these services is not particularly challenging thanks to the continuing advancing technologies. However, the guarantee of

high bandwidth performance of these services based-on a spontaneous on-demand pattern is fairly challenging. Moreover, providing high bandwidth with strict packet loss tolerances and high mobility is also difficult for the current networks which are commonly designed for best effort purposes.

Given that most Internet protocols are designed to comply with a best effort, or enhanced best effort paradigm, it is inevitable that the network will suffer from performance degradation in case of congestion. Recent work on deterministic networking (DetNet) aim to improve this situation by providing a ceiling on latency for a particular traffic flow, which significant improves packet error rate for specific DetNet services. This pioneering work gives a great example that new approaches are investigated to make the Internet aware of certain performance requirement other than the bandwidth.

Taking a look at the network infrastructure, service provider used to build dedicated network and resources for services requiring guaranteed performance. This is simply not cost-effective, neither is it flexible. The emergence of virtualization and VPN technologies make it possible to set up logically isolated computing and network instances from shared infrastructures. This can be used dedicatedly by specific services for improved performances. However, many questions are still to be answered as different technologies in various domains need to be combined to build network slices, which may require the separation of different resources and various types of performance guarantees.

2. Demand for Network Slicing

It is expected that a diversity of new services will emerge in 5G network. These services including smart home, industrial control, remote healthcare, Vehicle-to-Everything (V2X) and etc. will eventually create an ecosystem of "Internet of Everything". With hundreds of billions of devices from different business sectors connected, the future network needs to meet the diversified Quality of Experience (QoE) demands of different vertical industries. Typical QoE requirements for the end users or the applications are extremely low latency and high reliability, whilst the purchaser of the slice is looking for short time-to-market and rapid deployment of the service infrastructure needed to provide the technical underpinning of their business. Service providers' networks need to continuously evolve to adapt to this change. As a result, it is believed that future networks should be able to provide services with guaranteed performances together with the existing best-effort services. In order to achieve this, it is preferred that dedicated resources in the network could be used by different vertical industry

customers. Network slicing is proposed as an end-to-end solution for this purpose.

2.1. Guaranteed Service Performance

One of the most challenging requirements for future network is to provide guaranteed performance for varieties of new services whilst maintaining the economies of scale that accrue through resource sharing. It has been foreseen that the requirements of different services would be diversified and complex.

Taking augmented reality (AR) service as an example, it requires high bandwidth to provide a local video feed to the augmenter, and high quality augmented video back to the user. At the same time, it also requires extremely low latency since the created reality and the user's view must be synchronized to avoid reaction mismatch. Another example is the vehicular communications where the delay in traffic control system may directly jeopardize the road safety.

Network slicing can deal with these challenges by mapping the performance requirements to physically or logically dedicated resources.

2.2. End-to-end Customization

Customization is another significant feature of future services. Many vertical industries are expected to offer customization capabilities as a service to both internal manufacturing processes and specific end users. Meanwhile, these customized services need to be deployed with short time-to-market. The network needs to adapt to this challenge since customers may frequently adjust and refine their customization requirements.

There is ongoing work such as network orchestration, software defined networks and network function virtualization that aims to address this problem. In principle, these new technologies share a common request for the network to provide the ability to provide agile resource allocation.

2.3. Network Slicing as a Service

It is anticipated that the operation of 5G and future networks will involve new business models. Given that the network is more flexible, elastic, modularized and customized, the shared network infrastructure can be sliced and offered as a service to the customer. For instance, dedicated, isolated, end-to-end network resources with a customized topology can be provided as a network slice service to the tenant of this network slice.The tenants are

allowed to have a certain level of provisioning of their network slices.

3. Network Slicing Architecture

This section introduces the general system architecture of network slicing.

3.1. Basic Concepts

Network slicing is a collection of technologies that are used to establish logically dedicated resources including but not limited to connectivity, computing, storage, provisioning and specific network functions. The logical resources are a part of the larger common network infrastructures that are shared among various network slice instances. These dedicated resources can be customized to meet the diversified requirements of different vertical industries.

The following sections describe some basic concepts of network slicing.

3.1.1. Network Slicing Service Provider

A network slicing service provider, typically a telecommunication service provider, is the owner of the network infrastructures from which network slices are created. The network slicing service provider takes the responsibilities of managing and orchestrating corresponding resources that network slicing uses.

3.1.2. Network Slice Instance

A network slice instance (NSI) is the end-to-end realization of network slicing, which consists of the combination of physically or logically dedicated resources. An NSI typically associates with components from different network domains including core network, transport network and access network. It may also require cloud resources from data centres. Furthermore, end-user terminals may also allocate dedicated resource to a specific NSI.

Each NSI is defined and created for specific service-oriented requirements. The logically dedicated resources allocated to NSIs may be intrinsically isolated physical instances. They may also share common physical infrastructures according to implementation choices.

3.1.3. Network Slice Type

Network slices are categorized into different types according to the abstraction of characteristics of the services they facilitate. The methodology used for defining network slice types may be different for the owners of network slicing infrastructure. Some typical examples of network slice types according to 5G implementation include eMMB, mMTC and URLLC. Network slice type may be used to map specific network resources, VPNs, QoS categories according to real implementation. It is advised that mutual types should be defined according to existing main-stream service implementation scenarios. Extensions should be allowed for network slicing service provider to make according to new requirements.

3.1.4. Network Slice Template

A network slice template is an abstraction of the resource requirement for a set of similar network slice instances. Different templates are defined for individual network slice types. These templates are used to create certain network slice instances.

3.1.5. Network Slice Tenant

A network slice tenant is the user of specific NSIs, with which specific services can be provided to end customers. Network slice tenants can make requests of the creation of new network slice instances. Certain level of management capability should be exposed to network slice tenant from network slice service provider.

3.2. General Architecture

Figure 1 illustrates the general architecture of network slicing. It can be seen that two network slice instances are created from the shared network infrastructures. In principle, the network elements (NEs) represent any general network infrastructures for demonstration purposes. The two instances created do not know the existence of each other. However, they may share the computing, connectivity and storage resources of the NE, whether they are in physical or virtual forms. Meanwhile, the owner of a particular network slice instance is allowed to adjust the instance by requesting changes via the network slicing management and orchestration system.

```
+-----------------------------------------------------------+
|         Network Slice Management and Orchestration        |
|   +------------+ +-------------+ +--------------------+   |
|   |  Template  | |   E2E Slice | |   Life cycle Mngt. |   |
|   | Management | |Orchestration| |   and monitoring   |   |
|   +------------+ +-------------+ +--------------------+   |
|              Created Network Slice Instances              |
| +-------------------------------------------------------+ |
| |                                                       | |
| |  +---+          +---+            +---+                | |
| |  |NE1+----+     |NE3|            |NE5|                | |
| |  +---+    |     +-+-+            +-+-+                | |
| |         +-+-+     |                |                  | |
| |         |NE2+-----+                |                  | |
| |         +-+-+                      |  Network Slice   | |
| |           |                        |   Instance 1     | |
| |           +------------------------+                  | |
| +-------------------------------------------------------+ |
| +-------------------------------------------------------+ |
| |                                                       | |
| |  +---+                           +---+      +---+     | |
| |  |NE1+----+                   +--+NE5+------+NE6|     | |
| |  +---+    |                   |  +-+-+      +---+     | |
| |         +-+-+           +---+ |    |                  | |
| |         |NE2|           |NE4+-+    |                  | |
| |         +-+-+           +-+-+      |  Network Slice   | |
| |           |               |        |   Instance 2     | |
| |           +------------------------+                  | |
| +-------------------------------------------------------+ |
+-----------------------------------------------------------+

+-----------------------------------------------------------+
|             Physical Network Infrastructures              |
|    +---+         +---+             +---+      +---+       |
|    |NE1+----+    |NE3+------+   +--+NE5+------+NE6|       |
|    +---+    |    +-+-+      |   |  +-+-+      +---+       |
|           +-+-+    |      +-+-+ |    |                    |
|           |NE2+----+      |NE4+-+    |                    |
|           +-+-+           +-+-+      |                    |
|             |               |        |                    |
|             +------------------------+                    |
+-----------------------------------------------------------+
```

Figure 1. Network Slicing Architecture

It is fundamental to network slicing that slices may be created, the topology and/or its resources modified, and that the slices may be decommissioned in a timely manner with minimum work by the network slicing provider or the customer. This is not however unique to

network slicing, it is a goal of modern classical networks to be able to do this.

4. Data Plane of Network Slicing

In the network slicing architecture, the data plane in the edge and core of the network will likely be one or more of the standard IETF data planes: IPv4/IPv6, MPLS or Pseudowires (PW). This section assumes that the IETF protocol stack exists as-is, and describes the performance consideration in different layers of the data plane.

4.1. Propagation of Guarantees

Guarantees of delay start at the physical layer and propagate up the stack layer by layer. Any layer can add delay, and can take various steps to minimize the impact of delay on its layer, but no layer can reduce the delay introduced by a lower layer.

Guarantees of loss and jitter can, by contrast be upheld or improved at any layer of the protocol stack, but usually at a cost of increased delay. Where delay is a constrain as it is in some 5G applications the option of trading delay for better loss or jitter characteristics is not an option. In these circumstances it is critical that the quality characteristics start at the physical layer and be maintained at each layer of the protocol stack.

4.2. The Underlying Physical Layer

A point to point dedicated physical channel provides the delay, jitter and loss characteristics limited only by the media itself. This does not fulfil the need for rapid reconfiguration of the network to provision new services.

To address the need to provision a slice of the data-plane one approach that can be deployed is to time-slice access to the physical service. Ignoring many of the classic TDM offering as being too slow, a number of technologies are available that might be applied including OTN and FlexE. Whilst the provisioning of the channel provided by underlays such as FlexE and the interconnection of FlexE channels is within the scope of this architecture the operation of the underlay is outside its scope.

The logical sub-division of a physical channel be that a single channel with the full bandwidth available or a channel multiplexed at the physical layer such as is provided by FlexE we will consider in the following section.

4.3. Hard vs Soft Slicing in the Data-plane

Hard slicing refers to the provision of resources in such a way that they are dedicated to a specific NSI. Data-plane resources are provided in the data-plane through the allocation of a lambda, through the allocation of a time domain multiplexed resource such as a FlexE channel or through a service such as an MPLS hard-pipe. Note that although hard-pipes can be used to allocate dedicated, non-shared resources to an NSI, the using of allocation is bandwidth, which can result in more "lumpiness" in the physical channel that would not be present with a true physical layer multiplexing scheme.

Soft slicing refers to the provision of resources in such a way that whilst the slices are separated such that they cannot statically interfere with each other (one cannot receive the others packets or observe or interfere with the other's storage), they can interact dynamically (one may find the other is sending a packet just when it wants to, or the other may be using CPU cycles just when the other needs to process some information), which means they may compete for some particular resource at some specific time. Soft slicing is achieved through logically multiplexing the data-plane over a physical channel include various types of tunnel (IP or MPLS) or various types of pseudowire (again IP or MPLS). Although the design of deterministic networking techniques helps, it is not possible to achieve the same degree of isolation with these techniques as it is possible to achieve with pure physical layer multiplexing techniques. However where such techniques provide sufficient isolation their use leads to a network design that may be deployed on existing equipment designs and which can make unused bandwidth available to best effort traffic.

4.4. The Role of Deterministic Networking

Deterministic networking is a technology under development in the IETF that aims to both minimize congestion loss and set an upper bound on per hop latency. It allows a packet layer to emulate the behaviour of a fully partitioned underlay such might be provided through some physical layer multiplexing system such as FlexE.

Deterministic networking works by policing the ingress rate of a flow to an agreed maximum and then scheduling the transmission time of each flow to reduce the "lumpiness" and hence the possible buildup of queues and hence congestion loss.

Whilst deterministic networking is not as perfect as physical layer multiplexing in terms of latency minimization, because the scheduling is hop by hop and not end to end meaning that at each hop a packet has to wait for the transmission slot allocated to its flow, it has

the advantage that it is able to allocate slots not needed by the allocated traffic to best effort traffic. This reallocation of the unused transmission slots to background traffic significantly improves the efficiency of the network by amortizing the cost between the scheduled high priority users and the best effort users.

4.5. The Role of VPNs

VPNs are considered candidate technologies for network slicing. The existing VPN technologies mainly focus on the isolation of forwarding tables between different tenants and provide a virtual topology for the connectivity between different sites of a tenant. The VPN layer and the underlying network resources are usually loosely coupled, and statistical multiplexing is adopted to improve network utilization.

Although VPNs have been widely used to provide enterprise services in service provide networks, it is unclear that whether VPNs along with existing underlying tunnel technologies can meet the performance and isolation requirements of critical services in the vertical industries.

4.6. Dynamic Reprovisioning

A requirement of the network slicing system is that it can be dynamically and non-disruptively reprovisioned. That is not an unusual requirement of a modern network. However the frequency of reprovisioning with network slicing will be relatively high, such that it in many cases it is not possible to hide any disruption during a "quiet" time.

Physical multiplexing methods such as FlexE have the ability to seamlessly reprovision multiplex slots. At the network layer techniques such as make-before-break, segment routing, and loop-free-convergence can be used to provide uninterrupted operation during a topology change.

4.7. Non-IP Data Plane

Non-IP data plane in support of Information Centric Networking (ICN), some of the IoT services and other similar requirements will be added in a future version.

5. Control Plane of Network Slicing

There are two control plane systems that need to be considered. The first is the control plane of the slicing infrastructure itself, the second is the control plane of an individual slice.

The network slicing control plane receives instructions from the orchestration layer and creates the required network slices and manages them throughout their life cycle. The slices need to satisfy a diverse set requirements and need to be dynamically managed as the collective requirements of the set of network slices changes, and as the resource and capabilities of the physical network change with time. Changes occur as resources fail, and resources are added. They also occur as the slices are added and deleted possibly needing a garbage collection and defagmantation service.

The control plane of the network slicing system needs to comply with the SDN architecture, while still using distributed control protocols when it is necessary or proved to have advantages.

Within a slice the full range of existing control plane technologies needs to be permissible. Some slices will run the existing IGP protocols (such as IS-IS or OSPF) whilst others may use BGP. Some slices may be controlled by their own SDN controllers. However the architecture needs to be sufficiently general so as not to restrict the control protocols that may be used within a slice.

6. Management and Orchestration of Network Slicing

The management and orchestration layer of network slicing system is responsible for the slice template management, slice orchestration and life cycle management and monitoring of network slices. Network slice templates can be generated according to the functional and performance requirements of the tenants. In different network domains, different technologies may be used for network slicing, and orchestration is needed to build E2E network slice. The provisioning, runtime assurance and decommissioning of E2E network slices is also the key function of this layer.

It is expected that the management and orchestration layer would use state of the art management technologies to support short time-to-market, and help the operators to build an open ecosystem for new services in vertical industries.

7. Service Functions

To be provided in a future version.

8. OAM and Telemetry

To be provided in a future version.

9. IANA Considerations

This document makes no request of IANA.

10. Security Considerations

Each layer of the system has its own security requirements.

11. Acknowledgements

12. Normative References

[Network-Slice-White-Paper]
China Mobile Communication Corporation, Huawei Technologies Co. Deutsche Telekom AG,Volkswagen, "5G Service-Guaranteed Network Slicing White Paper", 2017, <http://labs.chinamobile.com/pdf/5GService-GuaranteedNetworkSlicingWhitePaper.pdf>.

[TD126_DraftRec_Y_IMT2020-NetSoft]
Nakao, A., Shimizu, T., and T. Kinoshita, "High level technical characteristics of network softwarization for IMT-2020", 2017.

[TD127_DrafSup_NetSoft-and-OSS_v0214-19H]
Goto, Y. and N. Morita, "Draft supplement to Y.IMT2020 series "Standardization and open source activities related to network softwarization of IMT-2020"", 2017.

Authors' Addresses

Liang Geng
China Mobile

Email: gengliang@chinamobile.com

Stewart Bryant
Huawei Technologies

Email: stewart.bryant@gmail.com

Jie Dong
Huawei Technologies

Email: jie.dong@huawei.com